América Negra & a Lei em Meados do Século XX

PENSAMENTO CRÍTICO SOBRE PESQUISA

Organizando o conteúdo

Montar este método foi bastante direto já que a nossa pergunta de pesquisa estava focada precisamente num único arquivo fonte e com uma janela de data de publicação claramente definida. O objetivo era usar o *Gale Digital Scholar Lab* para explorar e descobrir novas perguntas de pesquisa possíveis a partir de uma conjunto de documentos amplo, ao invés de investigar perguntas precisas em torno de um autor específico, gênero ou talvez outra variável. Dito isto, está claro que conforme trabalhamos com o conjunto de conteúdo, novas perguntas surgiram tornando necessário refinar o Conjunto de Conteúdo. Nós construímos duas novas versões do Conjunto de Conteúdo original, dividindo-o usando os valores de Tipos de Documento. Poderíamos ter feito isso já no início.

Outro caminho possível para Conjuntos de Conteúdo mais precisos - permitindo-nos explorar questões de pesquisa mais precisas - seria construir conjuntos de conteúdo adicionais com base na autoria, ou por caso. Tal precisão requer um conhecimento mais aprofundado e uma especialização cada vez maior no assunto em si. O *Gale Digital Scholar Lab* pode construir tais Conjuntos de Conteúdo, mas você, o pesquisador, precisa ter experiência suficiente no assunto a fim de definir os parâmetros de sua pergunta de pesquisa e como ela se relaciona com o Conjunto de Conteúdo que você venha a construir. Ter uma lista de casos envolvendo direitos civis ofereceria uma rica imagem dos pontos de vista dos afro-americanos durante esta Era. No entanto, significaria também determinar quais casos envolviam questões de direitos civis. Eram apenas aqueles que se ocupavam com os estatutos de direitos civis? Ou podemos aprender alguma coisa sobre como o sistema legal tratava ou via afro-americanos em casos que não estavam

diretamente relacionados aos estatutos de direitos civis? Estas são duas perguntas de tipos diferentes que requerem conjuntos de conteúdo diferentes.

Entendendo os Resultados

O que esses resultados nos dizem sobre nossos interesses de pesquisa? Podemos entender resultados de várias formas: como eles respondem as perguntas colocadas originalmente e como eles sugerem novas perguntas e novos caminhos de pesquisa.
Está claro que nossos resultados corroboram muitas coisas que já sabemos sobre a era dos direitos civis.

- Coincidiu com o auge da Guerra Fria e da ansiedade em torno do comunismo.
- Casos focados em discriminação por um lado e em igualdade por outro.
- Segregação escolar era uma preocupação primária dos casos que moldaram a experiência do sistema jurídico entre negros, pretos ou afro-americanos.

O que não está claro nos nossos resultados, no entanto, é se o sentimento dos documentos do nosso conjunto de conteúdo é devido ao seu gênero e contexto (ou seja, documentos são sempre negativos talvez porque envolvem contestação ou disputa) ou se está realmente relacionado a racismo e discriminação. Possivelmente é uma combinação das duas coisas, mas não tem nenhum método seguro de separá-los.
Também podemos examinar os resultados como uma forma de entender e pensar sobre a nossa metodologia. Que tipos de técnicas de construção de campos ou conjuntos de conteúdo poderíamos usar para criar coleções de documentos mais focadas ou precisas que pudessem responder melhor às nossas perguntas? Como poderíamos gerir se um documento realmente contém o que os metadados dizem que ele contém: em outras palavras, será que os conteúdos estão de acordo com o que os catalogadores e mesmo os autores originais, nos dizem? Uma leitura mais próxima dos documentos antes de adicioná-los a um conjunto de conteúdo nos ajudará a discernir se eles devem ser incluídos ou não.

Revisando Perguntas

Uma vez que vimos os resultados iniciais deste projeto utilizando o conjunto de conteúdo principal, ficou claro que poderíamos revisar nossas perguntas de pesquisa, tanto para deixá-las mais precisas, como também explorando uma nova pergunta em torno de algo inesperado, o bi-grama "partido comunista". De onde isso veio? Existe alguma forma de isolar isso para discernir porque aparece de forma tão proeminente no conjunto de conteúdo? Por que aparece em documentos que mencionam negros, pretos ou afro-americanos na metade do século XX? Existe uma sobreposição entre a Guerra Fria e o medo do comunismo e as tensões raciais da era dos direitos civis?

Como discutido no guia, não é só normal revisar suas perguntas de pesquisa depois de executar ferramentas de análise num conjunto de conteúdo, é uma parte integral do processo. Freqüentemente, a análise levantará novas questões que podem estar além do escopo de seu projeto atual. É assim que pesquisadores desenvolvem novos projetos e linhas de pesquisa, seguindo as pistas e novas perguntas que surgem enquanto realizam uma outra pesquisa.

Limitações do Projeto

Por mais útil que esses resultados possam ser, existem limitações quanto aos tipos de limpezas e análises que podem ser feitas com o *Gale Digital Scholar Lab*.

- Atualmente, não há nenhum método no *Gale Digital Scholar Lab* para comparar todas as palavras com um dicionário para identificar problemas de OCR. Iterações através das Configurações de Limpeza usando Modelagem de Tópicos é um método seguro para encontrar palavras problemáticas, mas é um processo que consome muito tempo. Substituições são fáceis de serem feitas, permitindo que OCR problemático seja consertado, mas não há método para encontrar todas as instâncias de palavras mal escritas.

- Este projeto não construiu conjuntos de conteúdos usando casos reais. Tampouco foram construídos através de uma leitura próxima dos documentos incluídos no conjunto. Um conjunto de conteúdo mais preciso poderia ser construído determinando, após o exame de cada documento, se seria apropriado ou não incluir num conjunto de conteúdo focado nos parametros específicos do projeto.

- Nós não consideramos completamente a diferença entre números puros de "contagens" e medidas estatísticas enquanto formas distintas de pensar sobre a importância. Embora o resultado da Modelagem de Tópicos nos permita examinar a contagem, o método de Alocação Latente de Dirichlet usado pelo MALLET é um tipo de previsão da probabilidade de que palavras apareçam juntas uma das outras. Sugere, em outras palavras, algo importante. Os N-gramas, em contraste, são contagens cruas através do conjunto de conteúdo. Ter mais documentos, ou documentos mais longos, com mais palavras, aumentaria essas contagens. No entanto, presença numérica nem sempre se traduz em importância intelectual ou sentido.

PARA ALÉM DO LAB

Apresentações

Todos os resultados das ferramentas podem ser baixados como imagens para serem usadas em PowerPoints ou inseridas em páginas web ou usadas de outras formas para apresentar seu trabalho.

Novas Visualizações

Também é possível fazer o download dos dados que alimentam as apresentações como separados por vírgulas (CSV) ou arquivos javascript object notation (JSON), permitindo que você crie e formate as suas próprias visualizações. Se você tiver a habilidade necessária, é possível colar ou criar novas visualizações que podem combinar resultados de visualizações similares em uma, permitindo que você faça a comparação e o contraste de formas novas que a ferramenta *Gale Digital Scholar Lab* não faz. Os downloads da ferramenta de Modelagem de Tópicos são especialmente ricos em possibilidades de nova visualização. O download da visualização de tópico é amplo e contém resultados para cada documento e medida para a ferramenta, muito mais dados do que as visualizações da Modelagem de Tópicos podem mostrar atualmente. Se você for um programador, este é o lugar ideal para começar a explorar dados criados pelo *Gale Digital Scholar Lab* usando outras ferramentas e designs de visualização.

Refinar conjuntos de conteúdo

Entender as limitações do *Gale Digital Scholar Lab* nos permite considerar o que pode ser feito tanto para construir conjuntos de conteúdo como para usar os resultados produzidos pelas ferramentas. Construir conjuntos de conteúdos usando os documentos e registros de casos da Suprema Corte dos EUA forneceria um método completamente diferente para considerar os temas de direitos civis e discriminação racial. Ao mesmo tempo, também isolaria a análise de documentos que estão explicitamente ligados a tais questões legais. No entanto, sabemos que esses temas não podem ser e não eram isolados em casos explícitos.

Projetos similares

Apreciar como podemos estruturar um projeto ou linha de investigação pode ser moldado tanto pelas ferramentas e conteúdo que temos, como pela modelagem de projetos similares que exploram tipos similares de documentos. O projeto Old Bailey Online (https://www.oldbaileyonline.org/) envolveu uma análise em larga escala dos processos judiciais do principal tribunal municipal da cidade de Londres. Embora seu conteúdo tenha sido codificado e limpo, e sua plataforma não contenha os mesmos tipos de ferramentas, como um projeto oferece uma forma de considerar como o exame dos registros da Suprema Corte dos EUA pode ser modelado. Ela fornece abordagens, perguntas e métodos que poderiam ser aplicados e testados usando o conteúdo e as ferramentas de nosso projeto atual.

PROJETO

América Negra & a Lei em Meados do Século XX

SINOPSE

O meio do século XX foi um tempo de imensa transformação para as pessoas de cor, em especial afro-americanos. Frequentemente citado como a era dos direitos civis, os anos 60 e início dos 70, assistiram a protestos, rebeliões, violência e eventualmente respostas legislativas às injustiças e discriminações raciais. Muitos homens e mulheres lutaram para mudar a forma como a sociedade americana tratava e entendia o lugar das pessoas de cor, seja nos ônibus, nas ruas, em casa, nas escolas ou no trabalho. Segregação era um dos principais pontos de disputa e o foco de um esforço legal considerável. Ao mesmo tempo, casos envolvendo afro-americanos avançaram através dos sistemas judiciais. Um exame dos Registros e Memoriais da Suprema Corte dos Estados Unidos revela o efeito das pressões dessa era no sistema jurídico assim como aquelas envolvidas em casos de direitos civis.

PERGUNTA CENTRAL DE PESQUISA

Como a linguagem sobre afro-americanos em documentos e arquivos jurídicos mudou entre 1950 e 1980?

Mais Perguntas Específicas:

- Os Registros e memoriais da Suprema Corte citam negros, pretos ou afro-americanos com um sentimento mais positivo ou negativo? Isso muda ao longo da era dos direitos civis?
- Quais são as frases ou colocações mais comuns usadas nos documentos mencionando negros, pretos ou afro-americanos? Quais são os tópicos que aparecem com maior frequência em documentos mencionando negros, pretos ou afro-americanos?
- Eles refletem temas que dominaram os conflitos na era dos direitos civis?

- Quais são as principais preocupações dos registros jurídicos mencionando negros, pretos ou afro-americanos durante esse período? Existem estados, estatutos ou outras entidades que se destaquem na análise?
- Existem diferenças entre Tipos de Documento no arquivo (Registros e Memoriais da Suprema Corte dos Estados Unidos)?

CONSTRUIR

Pesquisando

A pesquisa para este conjunto de conteúdos está limitada a um arquivo e a uma Data de Publicação definida.

Ferramentas específicas

Nenhuma das ferramentas precisou de conjuntos de conteúdo específicos.

Perguntas Específicas

A questão 6 poderia ser respondida utilizando o conjunto de conteúdo principal e assim foi necessário criar subconjuntos de conteúdo divididos por Tipos de Documentos: Memoriais e Petições; Declarações; Memorandos; Apêndices; etc.

LIMPAR

Foram necessárias diversas tentativas, ou iterações para acertar a limpeza em cada uma dessas análises. No final, foram necessárias várias Configurações de Limpeza diferentes, pois haviam diferentes palavras vazias necessárias para diferentes ferramentas, assim como diferentes abordagens de pontuação. Foi necessário acrescentar letras individuais à lista principal de palavras vazias (para cada letra, no caso de aparecerem como abreviações), assim como palavras vazias adicionais. Além disso, algumas substituições óbvias a partir do teste de Configurações de Limpeza. *Como as Configurações de Limpeza mudaram de acordo com as perguntas de pesquisa e a análise?*

Perguntas de pesquisa específicas:

Nenhuma configuração de limpeza adicional foi necessária para a Análise de Sentimentos.

Foram necessárias repetições consideráveis para remover as abreviações e letras individuais, assim como as palavras vazias adicionais.

Foram necessárias palavras vazias adicionais específicas para a Modelagem de Tópicos e que fossem diferentes das palavras usadas em N-gramas; “estado” e “corte”, por exemplo, precisaram ser mantidas em N-gramas (para “Estados Unidos”), mas removidas da Modelagem de Tópicos, que trata de elementos individuais. (Por exemplo, "Estados Unidos" nunca poderia ocorrer na Modelagem de Tópicos porque inclui duas palavras. Esse software não opera com frases.)

ANALISAR

Selecionar visões específicas para cada ferramenta foi muito direto. O tamanho do conjunto de conteúdos sugeriu uma busca que (1) procurou mais coisas; e (2) elevou a faixa para o corte dos resultados. Ambos por razões muito simples:

1. A Modelagem de Tópicos enquanto ferramenta discerne estatisticamente que palavras são mais propícias a aparecerem próximas umas das outras. Mais tópicos e mais palavras diminui o patamar do que é “significativo”, apresentando uma imagem mais refinada do que a análise estatística poderia sugerir. Em documentos muito similares como registros da corte, possivelmente teremos frases similares na medida em que perguntas e respostas são colocadas e decisões e argumentos são registrados.

1a. Selecionar mais palavras e mais tópicos é uma boa forma de peneirar essas similaridades “conhecidas” e podem trabalhar em conjunto com palavras vazias para ajudar a “penetrar” num grande conjunto de conteúdos.

2. Para N-gramas, uma abordagem similar ao pensar num possível “ruído” ajuda a buscar o equilíbrio entre o número de resultados e resultados significativos para o usuário. Existe um equilíbrio entre número e ruído.

Leia mais:

Modelagem de Tópicos
Pareceu melhor lançar uma rede mais ampla, em parte para ver que tipos de palavras apareciam nos modelos criados pelo software MALLET que alimenta a ferramenta. Necessitando mais palavras do que o padrão e duplicando os tópicos, produz tópicos mais refinados, refletindo o tamanho do conjunto de conteúdo. Esse exemplo utilizou 15 tópicos de palavras e 20 tópicos.

Análise de sentimentos
Esta ferramenta não possui configurações além da seleção a configuração de limpeza.

N-gramas
Assim como a Modelagem de Tópicos, parecia valer a pena ir além da configuração padrão de acordo com o tamanho do conjunto de conteúdo. Então, o limite para o número de vezes em que um N-grama tem que aparecer para ser considerado útil, foi elevado a quatro. Da mesma forma, o desejo de encontrar colocações, em vez de apenas palavras isoladas, levou a um ajuste do tamanho mínimo de N-grama para 2 (biGram) e do tamanho máximo para 5. Esse arranjo se traduz numa busca por "N-gramas de 2 a 5 palavras que aparecem no documento pelo menos 4 vezes ou mais".

Entendendo os resultados

Determinar sentido ou importância é uma parte crítica de uma investigação acadêmica. Um elemento essencial para isso num ambiente computacional de análise de texto como o *Gale Digital Scholar Lab* é entender que a contagem crua ou "mais resultados" nem sempre quer dizer algo importante; pode ser ruído, o que significa que está ali simplesmente porque o conjunto de conteúdo não passou por uma limpeza suficiente, ou porque não foram usadas as palavras vazias certas, ou talvez seja algo conhecido e esperado. Enfim, entender resultado realmente requer entender o que está sendo pedido das ferramentas nas configurações e parâmetros e como os resultados estão relacionados às variáveis selecionadas e aos algoritmos que alimentam a análise.

Modelagem de Tópicos

Refinar a lista de palavras vazias para a Modelagem de Tópicos, tomou algum tempo, pois a lista normal de palavras vazias não incluía palavras equivocadas do Reconhecimento Óptico de Caracteres (OCR). Este é um exemplo dos tipos de problemas que podem ocorrer, Tópico 6:

Revisão da Configuração de Limpeza para acrescentar "tihe", "andl", "anld", "that", "tlat", "tliat", "thie", "tlhat", "inl", removeu este tópico após a reexecução; produziu mais tópicos com palavras inúteis.

A Modelagem de Tópicos revela uma série de possíveis temas dentro dos Registros da Suprema Corte dos Estados Unidos:

> cidade, estado, público, polícia, brancos, peticionários, paz, alabama, lei, povo
> partido, comunista, comitê, governo, pontes, testemunho, membro, sindicato, atividades, membros
> condado, voto, eleição, votação, estado, distritos, cidade, população, eleitores
> estado político,queixosos, réus, ação, queixa, condado, faculdade de direito, escolas, conselho, educação, plano, estudantes, negros, brancos, altos, racial
> juri, grande, jurados, condado, negros, réus, estado, jurado, nomes, estado de julgamento, EUA, lei, direitos, lei, emenda, caso, federal, público, estatuto

Decidir quais as medidas de maior sentido ou importância nesses resultados pode ser complicado, dependendo de qual for a pergunta, é claro. Os resultados da Modelagem de Tópicos podem ser importantes simplesmente por serem inesperados ou novos. Ao mesmo tempo, eles podem confirmar algo que já se saiba, funcionando assim como uma referência para confirmar que o usuário está no caminho certo na leitura ou análise, ou em ambas. Como pode ser visto nos tópicos retornados acima, alguns são claramente relevantes para a era dos direitos civis. Alguns podem não ser. No entanto, há outras formas de entender esses resultados em relação ao conjunto de conteúdo.

O software que alimenta a Modelagem de Tópicos, MALLET, é especialmente bem conhecido e refinado. Oferece resultados muito ricos, que podem ser investigados de inúmeras formas olhando para a seção de Resultados de Tópicos, na visão por ferramentas e selecionando “Comparação de Tópicos”. Esta visão fornece uma lista de medidas descrevendo como os tópicos se relacionam ao conjunto de conteúdo e à análise.

Tokens: Esta métrica mede o número de palavras do conjunto de tópicos designadas para este tópico.

Entropia do documento: Esta métrica mede a probabilidade de que qualquer documento dado estará nos tópicos. Tópicos de baixa entropia virão de um pequeno conjunto de documentos enquanto tópicos de entropia mais alta virão de um conjunto de documentos mais amplo.

Comprimento padrão de palavras: Esta métrica mede o número médio de caracteres nos termos principais. Palavras mais longas são tidas como mais significativas, então maior comprimento de palavras indica mais tópicos específicos.

Coerência: Esta métrica mede o quão frequente as palavras no tópico aparecem próximas umas das outras. Quanto mais próximo de 0, mais provável de que os termos ocorram próximo uns dos outros.

Distância uniforme: Esta métrica mede a distância entre uma distribuição uniforme e a distribuição do tópico quanto as palavras atribuídas a ele. Quanto maior a distância, mais específico é o tópico.

Similaridade do Corpus: Esta métrica mede a distância entre a frequência das palavras no conjunto de conteúdo e a frequência das palavras atribuídas ao tópico. Quanto maior a distância, mais específico é o tópico.

Exclusividade: Esta métrica mede o quão exclusivo são os principais termos de cada tópico. Quanto mais alto o valor, maior a probabilidade de que os termos principais do tópico não apareçam como termos principais de outros tópicos.

Essas medidas permitem que os usuários explorem a forma como os tópicos criados pelo software se relacionam entre si e com o conjunto de conteúdo dos quais eles foram extraídos. O usuário pode olhar para a contagem pura, mas também todas as possíveis formas de inter-relação que as palavras, possuem entre si.

> Coerência se sobrepõe conceitualmente com N-gramas. Podemos comparar as duas ferramentas de forma significativa?
> Várias medidas apontam para especificidades ou, para colocar de outra forma, para a precisão ou clareza de um tópico dentro de um conjunto de conteúdo e documentos:

Entropia do Documento e Distância Uniforme oferecem formas de examinar a maneira como tópicos específicos se encaixam no conjunto de conteúdo, assim como em documentos. Outra pergunta está relacionada a singularidade de um tópico. Similaridade do Corpus oferece meios para pensar sobre a excepcionalidade de um tópico.

Nomes para os tópicos criados pela ferramenta neste exemplo são fornecidos para que possam servir como referência depois. Esses nomes aparecerão na visão de Proporção de Tópico, substituindo "Tópico [número]", tornando mais fácil a navegação nos resultados.

A ferramenta de Modelagem de tópicos nos permite avançar através dessas medidas e dos próprios tópicos através da visão de Proporção de Tópico. Podemos selecionar documentos específicos por título e comparar os tópicos que aparecem. Isso é especialmente útil como forma de penetrar no próprio conjunto de conteúdo.

Nos nossos resultados vemos que os tópicos com maior presença na visão de Proporção em cada caso são aqueles relacionados a termos jurídicos ou processuais. Isso não é surpreendente, mas também não é particularmente útil.

Análise de Sentimentos

Este resultado parece sem surpresas, considerando o que sabemos sobre a era dos direitos civis, um tempo quando manifestações não-violentas eram respondidas com violência física e a discriminação racial estava incorporada no cerne da infraestrutura socioeconômica nos Estados Unidos. No entanto, desperta **NOVAS** perguntas de pesquisa:

1. Seria possível associar ou conectar os anos com pontuação de sentimento mais baixas a eventos o momentos específicos da era de direitos civis?
2. Porque talvez haja anos com pontuação positiva ao longo dos anos 70?
3. Até que ponto é possível atribuir essas pontuações de sentimento às questões que moldaram a era dos direitos civis versus a natureza geral de casos jurídicos como contestações ou conflitos, ou questões de atrito? Por exemplo, será que a maioria dos Registros da Suprema Corte dos Estados Unidos possuem um sentimento negativo porque lidam com questões jurídicas ou existe uma tendência racial implícita nos registros que tratam de americanos negros?

A pergunta final é especialmente importante para considerar de acordo com o que sabemos sobre as armadilhas da Análise de Sentimentos. Análise de Sentimentos fomenta um léxico para informar as pontuações atribuídas a cada palavra com conjunto de conteúdo. Nós precisamos considerar as palavras e pontuações no léxico para obter uma compreensão abrangente da visualização resultante. Esse contexto é imperativo quando consideramos as implicações raciais do conjunto de conteúdos. Por exemplo, palavras como negros, afro-americanos, estão incluídas no léxico? Se estão, quais são as pontuações de sentimento associadas? Como isso se traduz na visualização? As respostas a essas perguntas informam a nossa análise crítica dos resultados.

Leitura recomendada:

Rice, D., Rhodes, J. H., & Nteta, T. (2019). Racial bias in legal language. Research & Politics. https://doi.org/10.1177/2053168019848930

N-gramas

Configuração: Min 2 Max 5, Limite 5

Limpeza: Jurídico Estados Unidos Sem pontuação Sem números

Esses são os N-gramas principais, após reexecutar a ferramenta diversas vezes com diferentes Configurações de Limpeza. Estranhamente, o bigrama mais frequente (N-grama com duas palavras ou tokens), é “partido comunista”. Os poucos que estão a seguir, no entanto, lidam diretamente com temas esperados: "direitos civis", "conselho educação" (possivelmente para "conselho de educação”, "emenda quatorze", "conselho escolar", "igual proteção", etc. Tudo isso diz respeito às principais batalhas legais em torno das questões de raça na era dos direitos civis dos anos 60 e início dos anos 70.

Também sugere que as questões mais frequentes que a Suprema Corte dos Estados Unidos tratou em relação à era dos direitos civis tiveram a ver com a escolarização e o acesso a ela, e à segregação. No entanto, apesar do fato de que a discriminação racial era o cerne da questão, ela vem muito abaixo na lista, seguindo a terminologia de igualdade de direitos. Isto sugere que, embora os apelantes estivessem bem cientes da discriminação racial, eles buscaram proteção legal usando argumentos de igualdade de direitos, em vez de se concentrarem na discriminação, como base para seus processos legais. É importante notar que a busca original não se relacionou especificamente a "direitos civis" ou "discriminação"; estes termos apareceram como uma questão de

análise.

Tais resultados confirmam muitas coisas já conhecidas sobre a era dos direitos civis e os procedimentos legais. O destaque do termo "partido comunista", entretanto, sugere uma possível nova direção para a pesquisa.

1. Porque o termo ”partido comunista” aparece de forma tão proeminente em textos mencionando Americanos Negros nos registros da Suprema Corte entre 1950 e 1980?
2. Que conexões podem ser feitas entre a Guerra Fria e raça na luta por direitos civis em meados do século XX nos Estados Unidos?

EXPANDINDO O PROJETO

Leia mais sobre formas como você pode expandir esse projeto com iterações, perguntas de pesquisa e análises.

Leia mais

Iteração

A iteração para este projeto focou mais na Configuração de Limpeza do que em trabalhar no conteúdo em si. Isto é o comum; a Configuração de Limpeza padrão é só um ponto de partida. Cada projeto precisará ter pelo menos uma Configuração de Limpeza própria, ou mais, que vai requerer testes e reexecução até que os resultados da análise sejam significativos e livres de dados "cheios de ruído".

Ao longo do curso deste projeto, se tornou cada vez mais claro que embora o conjunto de conteúdo fosse utilizável para a Análise de Sentimentos sem muitos ajustes na Configuração de Limpeza, N-gramas e especialmente a Modelagem de Tópicos precisou precisaram de alguns ajustes para chegar à Configuração de Limpeza correta. Isso tem tanto a ver com a forma como as ferramentas funcionam como com OCR problemático. A Análise de Sentimentos combina palavras que já conhece, portanto se algo estiver escrito errado, é ignorado. Limpeza, nesse contexto, só adiciona palavras à mistura; OCR problemático não é analisado. As outras duas ferramentas trabalham com as palavras que estão efetivamente num documento, então se palavras com problema de OCR têm uma presença suficientemente significativa, irão aparecer como outras palavras.

1. No caso das N-gramas, teve N-gramas que incluíram palavras escritas corretamente, mas que tiveram pouca utilidade para a pergunta de pesquisa. Por

exemplo, os números e abreviações de seção que fazem parte da maioria dos documentos jurídicos, apareceram. Quando foram adicionados a lista de palavras vazias, deixaram de ser incluídos na análise de N-grama. Foram necessárias várias tentativas para encontrar a todos, mas foi possível obter um resultado razoavelmente limpo e significativo depois de algumas iterações.

2. Modelagem de Tópicos, no entanto, demorou muito mais, pois a natureza da ferramenta é de recolher palavras que estatisticamente aparecem juntas com frequência. Enquanto OCR problemático foi geralmente ignorada por ferramentas de N-grama, rapidamente se tornou aparente com a Modelagem de Tópicos porque a ferramenta produziu resultados sugerindo que palavras OCR problemáticas em si constituíam um ou mais tópicos. Reexecutar a análise de Modelagem de Tópicos, permitiu que o resultado a ser usado revisasse a Configuração de Limpeza; e aí a análise foi executada de novo.

2a. Outro elemento problemático com a Modelagem de Tópicos é encontrar o equilíbrio entre encontrar resultados alinhados com o gênero do documento e eliminar palavras através da lista de palavras vazias para assegurar que esses resultados são significativos. Na visão de Proporção de Tópicos, os tópicos prevalecentes eram aqueles relativos a termos legais ou processuais, não importa quantas tentativas foram feitas para limpar OCR ou palavras secundárias (tais como preposições, conjunções, artigos, etc.) Mas remover termos como "estatuto" ou "federal" talvez altere temas importantes para a busca ainda que apareçam com frequência como tópicos unificados. Isso seria esperado, mas é um outro tipo de "ruído legítimo" - resultados que devem ser percorridos a fim de encontrar os temas críticos da era dos direitos civis.

Resultados de pesquisa

Está claro através deste curto projeto-teste que a análise em larga escala dos Registros da Suprema Corte dos Estados Unidos fornecem perspectivas sobre temas e preocupações amplas que se poderia imaginar que seriam encontradas nos procedimentos legais da era dos direitos civis que mencionam negros, pretos ou afro-americanos. Ao mesmo tempo, os resultados também sugerem novas questões a serem consideradas.

Questões originais

1. Os Registros e Memoriais Processuais da Suprema Corte dos Estados

Unidos possuem claramente um tom um tanto negativo. Até quando dividido por Tipo de Documento, o sentimento negativo é impactante. Flutua, e fora dos Memoriais e Petições, tende a um tom ligeiramente mais positivo ao longo dos anos 70.

2. "Partido comunista", "direitos civis", "conselho educação", "emenda 14", "conselho escolar", "igual proteção", "grande júri", "ensino médio", "procurador geral", "lei de direitos", "lei de direitos civis", "discriminação racial", "raça cor", "título vii", "direitos constitucionais"
3. Isso requer uma Configuração de Limpeza mais precisa para remover OCR problemáticos. Alguns tópicos no entanto refletem temas de direitos civis, em particular segregação e discriminação, assim como direitos iguais.
4. Isso não é fácil de discernir dos nossos resultados. Modelagem de Tópicos sugere algumas possibilidades, mas precisa ser mais clara. N-gramas também sugere temas dominantes, mas não são tão explícitos como poderiam.
5. No N-gramas os termos "emenda quatorze" e "título vii" ambos proíbem expressamente a discriminação baseada em raça – o primeiro como parte da constituição e o último como uma sessão da Lei de Direitos Civis de 1964.
6. Parece que, especialmente se tratando de N-gramas, é necessária uma exploração maior.

Novas perguntas

1. Por que "partido comunista" aparece no N-gramas?
2. De que forma o "Tipo de Documento" afeta os resultados das ferramentas de análise?

Revisando o Conjunto de Conteúdo

À luz dos resultados iniciais, faz sentido revisar ou subdividir o conjunto de conteúdo original para atender as novas perguntas de pesquisa. Um dos meios possíveis para isso é criar subconjuntos de conteúdo a partir do conjunto de conteúdo original usando vários metadados, tais como Tipos de Documentos. Por exemplo, um dos Tipos de Documentos no *Gale Digital Scholar Lab* são os memoriais e petições, que constituem pedidos e processos apresentados à Suprema

Corte dos Estados Unidos.

Talvez a primeira questão sobre "partido comunista" possa ser esclarecida através da criação de subconjuntos de conteúdo compostos por diferentes
Tipos de Documentos. A criação de um conjunto de conteúdos que contenha apenas memoriais e petições, e um segundo com o restante dos Tipos de Documentos, pode fornecer uma visão diferente da questão da pesquisa, contextualizando todas as análises que têm sido realizadas com a questão do "gênero". O gênero permite certos tipos de perguntas que podem moldar a forma como se pensa sobre os resultados das ferramentas:

1. O que é um “memorial” ou uma “petição”? São a mesma coisa? Quem os escreve e por quê?
2. Que tipo de informações estão contidas nos memoriais e petições?Enquanto um gênero, possui características em especial?
3. De que forma o conhecimento sobre um Tipo de Documento, um gênero ou uma forma de escrever, molda nossas perguntas de pesquisa? O que podemos aprender ao entender a forma de um documento e o que ele pode conter textualmente?

Subconjunto de Conteúdo: Memoriais e Petições Modelagem de Tópicos

> Júri estado julgamento grande EU condado solicitante negros morte jurado discriminação EU título minoria racial emprego vii preto cir programa escola negro raça branco escolas estado educação público lei igual escola escolas conselho plano educação racial estudantes negros dessegregação condado cidade habitação propriedade pública privada estado EU parque discriminação racial

Este subconjunto de conteúdo parece estar mais precisamente preocupado com as questões da era dos direitos civis do que o conjunto principal.

Análise de Sentimento

Parece que Memoriais e Petições permanecem consistentemente negativos em termos de análise de sentimentos, exceto por alguns poucos anos positivos em meados dos anos 70. Isso poderia sugerir que, enquanto um gênero, Memoriais e Petições possui geralmente um tom e sentimentos negativos e que talvez a negatividade não esteja associada a questões da era dos direitos civis. Ou talvez seja exatamente o oposto. Uma boa forma de checar seria executar a mesma Análise de

Sentimentos num novo Conjunto de Conteúdo contendo documentos que não são memoriais ou petições no conjunto de conteúdo principal.

N-gramas

Podemos ver que Partido Comunista não aparece mais como principal N-grama, e que os mais ou menos 10 principais N-gramas restantes estão focados exclusivamente em questões da era dos direitos civis. O Partido Comunista aparece como 12o N-grama, numa queda dramática do primeiro lugar. Claramente ainda guarda certa relevância, mas não tão proeminente como no conjunto de conteúdo principal que incluía memorandos.

Subconjunto de Conteúdo: Declarações, Memorandos, etc.

Este é o resultado das mesmas análise com outros Tipos de Documento.

Modelagem de Tópicos

> Escola escolas conselho plano educação estudantes alto preto branco crianças
> partido comunista nacional exibição mcgohey guerra sacher político objeção classe
> partido comunista pontes testemunho harry
> comitê união resposta membro donohue estado
> defensores fol defensor demandantes alabama
> condado reclamante moção reclamação

Tem alguns tópicos claramente relacionados a era dos direitos civis, mas também alguns focados no tópico do Partido Comunista, assim como outros temas. Isso sugere uma diversidade maior de conteúdo neste subconjunto de conteúdo do que no outro.

Análise de Sentimento

Análise de Sentimento está dramaticamente diferente. Declarações e memorandos claramente tinham tons bem diferentes em termos de raça e após os anos 60. Mas o cerne da era dos direitos civis dos anos 60 continua com um sentimento majoritariamente negativo.

N-gramas

Note-se que "partido comunista" ainda aparece dramaticamente em primeiro lugar.

Estava em 12o nos Memoriais e Petições. Ainda assim aparecem muitos dos mesmos N-gramas, sugerindo que questões da era de direitos civis permanecem dominantes ao longo dos documentos independente do tipo.

REFLEXÕES SOBRE O MÉTODO

Este projeto fornece perspectivas sobre como montar conjuntos de conteúdo e o que significa a iteração quando são revisados e limpos antes das análises. As comparações entre os dois subconjuntos de conteúdo, Memoriais e Petições e Declarações, Memorandos, etc, nos permitem ver como que os parâmetros ou campos que são usados para construir um conjunto de conteúdo podem afetar ou moldar os tipos de resultados obtidos das ferramentas de análise. Aparentemente, "partido comunista" foi assunto de discussão em memorandos e não em memoriais ou petições da Suprema Corte dos EUA entre 1950 e 1980 em casos envolvendo menções a negros, pretos ou afro-americanos.